

CLAIMS

5u7 A127

1. A process for determining latency between multiple servers and a client
 5 across a network in a computer environment, comprising the steps of:
 receiving a request for latency metrics on a content server;
 wherein said latency metric request specifies a particular client;
 providing a latency management table;
 wherein said latency management table comprises a list of IP addresses
 10 along with corresponding Border Gateway Protocol (BGP) hop counts, dynamic
 hop counts, and Round Trip Times (RTT);
 looking up the latency metric for said client in said latency management
 table;
 sending said latency metric to the requesting server;
 15 wherein the BGP hop count for said client in said latency management
 table is used for said latency metric upon the first request for said client; and
 wherein the dynamic hop count and RTT data for said client in said
 latency management table are used for said latency metric for subsequent
 requests for said client.
 20
2. The process of Claim 1, further comprising the steps of:
 sending periodic latency probes to the IP addresses in said latency
 management table;
 receiving response packets for said latency probes; and
 25 recording the dynamic hop count and latency (RTT) data in said latency
 management table.
3. The process of Claim 2, wherein periodic latency probes are sent to a
 higher level server of a client by masking said client's IP address in said latency
 30 management table.

5

10

15

20

management table;

25

table is used for said latency metric upon the first request for said client; and

30

Sub A12 7

7. The apparatus of Claim 6, further comprising:
a module for sending periodic latency probes to the IP addresses in said latency management table;

5 a module for receiving response packets for said latency probes; and
a module for recording the dynamic hop count and latency (RTT) data in said latency management table.

8. The apparatus of Claim 7, wherein periodic latency probes are sent to a
10 higher level server of a client by masking said client's IP address in said latency management table.

9. The apparatus of Claim 7, further comprising:
a module for receiving requests for a content server address from said
15 client;

a module for sending a latency metric request to the appropriate content servers;

a module for receiving latency metric data from said content servers;
a module for determining the optimal content server for said client; and
20 a module for sending said optimal content server's address to said client.

10. The apparatus of Claim 9, wherein said determining module gathers the expected latency metrics and uses the inverse relationship of the hop counts in said latency metric data in a weighted combination with the RTT in said latency
25 metric data to determine which latency metric data indicates the optimal content server.

11. A program storage medium readable by a computer, tangibly embodying a program of instructions executable by the computer to perform method steps
30 for determining latency between multiple servers and a client across a network in a computer environment, comprising the steps of:

Sub A12-7

receiving a request for latency metrics on a content server;
wherein said latency metric request specifies a particular client;
providing a latency management table;

wherein said latency management table comprises a list of IP addresses
5 along with corresponding Border Gateway Protocol (BGP) hop counts, dynamic
hop counts, and Round Trip Times (RTT);

looking up the latency metric for said client in said latency management
table;

sending said latency metric to the requesting server;

10 wherein the BGP hop count for said client in said latency management
table is used for said latency metric upon the first request for said client; and

wherein the dynamic hop count and RTT data for said client in said
latency management table are used for said latency metric for subsequent
requests for said client.

15 12. The method of Claim 11, further comprising the steps of:

sending periodic latency probes to the IP addresses in said latency
management table;

receiving response packets for said latency probes; and

20 recording the dynamic hop count and latency (RTT) data in said latency
management table.

13. The method of Claim 12, wherein periodic latency probes are sent to a
higher level server of a client by masking said client's IP address in said latency
25 management table.

The method of Claim 1, comprising:

- receiving requests for a service;
- sending a latency metric to a client;
- receiving latency metric data from the client;
- determining the optimal service based on the latency metric data;
- sending said optimal service to the client.

The method of Claim 1, comprising:

- receiving latency metrics from a plurality of clients;
- receiving latency metric data from the clients;
- determining the optimal service based on the latency metric data;
- sending said optimal service to the clients.

- 5

- 10

Q&A